Race to the Top Assessment Program
Written Input
Submitted before November 8, 2009

Author	Title	Date Submitted
Elaine Hackett	No Title	10/22/2009
William Damour	No Title	10/23/2009
Mark Kasmer	No Title	10/26/2009
Paul Nichols	A Framework for Evaluating and Planning Assessments Intended to Improve Student Achievement	11/06/2009

[redacted]

From: Elaine Hackett [ehackett2@cinci.rr.com]
Sent: Thursday, October 22, 2009 2:54 PM
To: Duncan, Arne
Subject: re: new tests for new standards??????

Dear Secretary Duncan,

I attended elementary school during the fifties, taught during the sixties and taught again, during these beginning years of the new millenium. Why can't the standardized achievement tests that have been used for years (with changes over the years) and have proven to be excellent indicators of achievement be used??? Why do we have to continue re-inventing the "wheel." Every state has written new standards and yet there is supposedly a set of "national standards." (50 sets when there really only needs to be one set)

Iowa Tests of Basic Skills, Terra Nova Tests, and Stanford Achievement Tests have been used for many years with excellent results of measuring student achievement. I'm assuming these tests "test" the achievements reflected by the standards that have been created over the past 10 years both nationally and state-by-state. Why reinvent this wheel??

When I returned to teaching in 2001, after being away for 30 years, I was surprised that curriculum guides were no longer in existence (at least not in Ohio.) The guides guided me through the 60s (albeit, in Massachusetts), but apparently faded away during the 30 years I was away. In 2000, I was handed three (3) 3" binders full of words, that supposedly were "a curriculum guide." Lots of words.

Anyway, I just read that the DOE was seeking advice from testing experts and the public by traveling across the country seeking ideas about what should be tested if there is a national set of standards. Why not use the above mentioned tests to match the national standards and save lots of money?

We are worried about how America ranks among industrialized and developing nations as far as education is concerned. We are one nation (that just happens to be divided into 50 states) that should have "one" set of standards and one test to determine the achievements of our students as they learn the indicators associated with each of the standards. Much money would be saved, the educational bureaucracies that exist across our nation could be significantly reduced, and the students and teachers would benefit tremendously.

I have other comments about what is or is not going on in the realm of Education, but I think I'll save my thoughts for another email.

Thank you.

Sincerely yours,
Elaine Hackett
Licensed Ohio teacher - Kindergarten through Grade Eight, including Middle School Math

P.S. I do consider myself "Highly Qualified" - I know my subject(s) and I enjoy working with children, particularly those on the lower end of the Bell Curve

P.S.S. I notice there is no mention of enjoying working with children in the definition of a "Highly Qualified" teacher. (A person can know their major extremely well, be extremely intelligent, pass all the tests presented to them BUT not enjoy children)

[redacted]

From: William Damour [William.Damour@mpls.k12.mn.us]
Sent: Friday, October 23, 2009 10:59 AM
To: Race To The Top Assessment Input
Subject: Race To The Top Assessment Program

Please review the SCANS Commission work. Those skills are still very relevant to a broad rigorous and relevant education.

[redacted]

From: Mark Kasmer [mark.kasmer@staff.spartaschools.org]
Sent: Monday, October 26, 2009 10:08 AM
To: Race To The Top Assessment Input
Subject: ELL Testing

Good Afternoon, I read a recent Blog concerning ELL testing and related concerns w/ NCLB requirements. I am an Assistant Supt. of Sparta Area Schools, located outside Grand Rapids in the state of Michigan. We have any where from 150 to 300 ELL /migrant students each year in a district just shy of 3,000 students. The students have to take a English Proficiency Screener upon arrival in our district, usually the fall. Additionally, the students are given an English proficiency test in the Spring. Finally, the MEAP or state test is administered in the fall grades 3-9 and MME/Act in the Spring for HS students.

If a student is a migrant student, who are often ELL, then they are here from Sept-October, during which time they are administered the screener and the MEAP. Consequently, out of the 6 weeks of time spent in the district, more than ½ of the time is spent on testing and/ or dealing with the effects of limited instructional as a result of staffing adjustments so to fit the required testing schedules. Since they are not here in the Spring, the migrants are re-issued the screener, so to adhere to the requirement of annual English development testing. This test is really not developed for this purpose, so the state is considering and additional test for fall English language development of migrant students. This is totally unproductive time for the students and staff. We need to revise the requirements for evaluating students to fit the needs of the learners 'schedules.

I understand the need to monitor progress, yet we should be spending the limited time we have w/ migrants on instruction, not assessing. I suspect this is a National issue, and should have a National resolve. If students were to be evaluated w/ a common English development test w/ common assessment dates across the country, then at least they would only have one test for English language development. It would not matter where they were at the time. Additionally, perhaps these students should forgo the state tests, at least until they test-out or meet a minimum English language development score. Why test them on a state exam, when we know they do not have the language skills to receive an accurate score?

I am sure there are many considerations for testing ELL students, yet the testing is not the product. The intent is to educate students. We are certainly spending an inordinate amount of time testing which has little bearing on instructional supports; rather. compliance. I do not claim to be an expert in this field and perhaps my recommendations are a bit naïve, yet I do know we are going to great lengths to meet the federal mandates in assessing , more so than focusing on the instructional needs of the ELL learner, especially those who are from seasonal or migrant families. I am happy to address further w/ more detail should you desire additional input.

Thank-you for your consideration, Mark

Educational Measurement: Issues and Practice
Fall 2009, Vol. 28, No. 3, pp. 14–23

A Framework for Evaluating and Planning Assessments Intended to Improve Student Achievement

Paul D. Nichols, Jason L. Meyers, and Kelly S. Burling, *Pearson*

Assessments labeled as formative have been offered as a means to improve student achievement. But labels can be a powerful way to miscommunicate. For an assessment use to be appropriately labeled "formative," both empirical evidence and reasoned arguments must be offered to support the claim that improvements in student achievement can be linked to the use of assessment information. Our goal in this article is to support the construction of such an argument by offering a framework within which to consider evidence-based claims that assessment information can be used to improve student achievement. We describe this framework and then illustrate its use with an example of one-on-one tutoring. Finally, we explore the framework's implications for understanding when the use of assessment information is likely to improve student achievement and for advising test developers on how to develop assessments that are intended to offer information that can be used to improve student achievement.

Keywords: formative assessment, validity, diagnostic assessment

The assessment and accountability provisions of No Child Left Behind (NCLB; Public Law 107–110, 2001) set targets for student achievement (Baker, 2004). In part in response to these targets, educators have set about exploring means of improving student achievement. Assessments labeled as formative have been offered as a means to customize instruction to narrow the gap between students' current state of achievement and the targeted state of achievement.

But labels can be a powerful way to miscommunicate. As Messick (1980, 1981, 1989) argues, when a test is labeled, it is judged (if only tacitly) as belonging to the broader category represented by that label. Users tend to ascribe all of the values and attributes associated with that category to the test. When an assessment is labeled formative, what appears to happen is "operationism in reverse" (Coombs, Raifa, & Thrall, 1954). The assessment that has been labeled a formative assessment is stereotyped and endowed with all the values and attributes commonly associated with formative assessments. The assessment is assumed to be capable of providing educators the information needed to improve student achievement. In the mind of the consumer, this distinction is bestowed upon the assessment without the provision of any theoretical rationales or empirical evidence to warrant such a gift.

The label "formative" is applied incorrectly when used as a label for an assessment instrument (Stiggins, 2001; Wiliam, 2006; Wiliam & Black, 1996). In technical discussions, the use of the phrase "formative assessment" is an implied claim of validity. Just as validity refers to a particular interpretation and use of assessment scores, reference to an assessment as formative is shorthand for the particular use of assessment information, whether coming from a formal assessment or teachers' observations, to improve student achievement. As Wiliam and Black (1996) note: "To sum up, in order to serve a formative function, an assessment must yield evidence that, with appropriate construct-referenced interpretations, indicates the existence of a gap between actual and desired levels of performance, and suggests actions that are in fact successful in closing the gap" (p. 543).

For an assessment use to be appropriately labeled "formative," both empirical evidence and reasoned arguments must be offered to support the claim that improvements in student achievement can be linked to the use of assessment information by an instructional agent such as a teacher, instructional software, or the learners themselves. But marshaling evidence and constructing arguments to support

Paul D. Nichols is Vice President at Pearson, 2510 North Dodge Street, Iowa City, Iowa 52240; paul.nichols@pearson.com. Jason L. Meyers is a Research Scientist at Pearson, 400 Center Ridge Drive, Austin, Texas 78753. Kelly S. Burling is a Research Scientist at Pearson, 3502 Westover Road, Durham, NC 27707.

the claim that an assessment use may appropriately be labeled "formative" involves more than evidence of student gains that coincide with assessment administration. Rather, the argument must causally link information from performance on a particular assessment to the selection of instructional actions whose implementation leads to gains in student learning.

Establishing a causal link between information from performance on a particular assessment to gains in student learning has proved difficult. For example, Black and Wiliam (1998) argued for a causal link between classroom-based formative assessment and students' gains in achievement. Yet throughout a series of eight examples alternative explanations continually threatened the internal validity of the conclusion that information provided by classroom assessment caused students' achievement gains. As Black and Wiliam (1998) note: "The examples do exhibit part of the variety of ways in which enhanced formative work can be embedded in new modes of pedagogy. In particular, it can be a salient and explicit feature of an innovation, or an adjunct to some different and larger scale movement—such as mastery learning. In both cases it might be difficult to separate out the particular contribution of the formative feedback to any learning gains" (p. 16).

Our goal in this article is to support the construction of such an argument by offering a framework within which to consider evidence-based claims that information from performance on a particular assessment can be used within specified contexts to improve student achievement. This framework is an elaboration of the more general framework offered by Messick (1989) and further developed by Kane (2001, 2006). In the first section, we describe this framework. Next, we illustrate the use of this framework with an example of one-on-one tutoring. Finally, we explore implications of this framework for understanding when the use of assessment information is likely to improve student achievement and for advising test developers on how to develop assessments that are intended to offer information that can be used to improve student achievement.

Framework

General validity theory appears to privilege test score interpretation over test score use. Controversy exists even on the acceptability of test score use consequences as validity evidence (Green, 1998; Mehrens, 1997; Reckase, 1998). In contrast, validity claims with regard to formative assessment emphasize test score use over test score interpretation. The consequences of formative assessment use, in terms of improved student learning, are arguably accorded more importance than other sources of evidence. The claim for formative assessment is that the information derived from students' assessment performance can be used to improve student achievement. It is how that information is used, not what the assessment tells us about current achievement, that impacts future achievement. Therefore, use, based on a valid interpretation, is the primary focus of the validity argument for formative assessments.

The emphasis on test score use to improve student achievement broadens the focus of validity investigation to include the system within which test score information is employed. The evaluation of a formative claim must be done within a systemic framework rather than treating assessment scores in isolation. Note that as we discuss this system we shift from referring to tests and test score information to assessments and assessment information. This shift reflects the understanding that assessment results include quantitative scores, such as ability estimates, as well as qualitative judgments such as teachers' appraisals.

The argument that information from assessment scores may be used to improve student achievement implies a certain framework for evaluating the validity of these claims. The framework offered here is an elaboration of the more general validity theory (Kane, 2001, 2006; Messick, 1989) and may be characterized as an interpretive argument (Kane, 1992, 2006) for the formative use of test scores. Formative assessment information is represented as a component of a system of coordinated assessment and instruction that eventually leads to improvements in student achievement. This system consists of a number of components and a sequence of interpretations, as represented by the framework in Figure 1. The framework shown in Figure 1 is neither complete nor exhaustive; however, it does serve to illustrate the complex reasoning that must support any claim that the information offered in student performance data can be used to improve student achievement.

The framework in Figure 1 represents a flow of activities grouped into phases. The first two phases comprise the *formative system*. The initial phase of activities has been labeled the *assessment phase*. In this phase, information intended to be used to improve student achievement is extracted from student behavior. Ideally, this information is used to prescribe appropriate instruction in the *instructional phase*. The instruction is appropriately implemented and leads to student learning. Finally, a *summative phase* of activities follows the formative system. Student behavior is again observed but now student behavior is used to extract summative information as evidence relevant to the formative claim of the formative system.

Framework Components

This framework comprises two kinds of components: structures and actions. The structures are represented in the figure as rectangles. Structures represent organized information and may be static or dynamic. For example, structures may be sets of data. The system within which assessment information is employed is data-driven. The student data component often comprises 0 and 1 values in a student by item matrix. Alternatively, the structures may be content frameworks that define what students are expected to learn or instructional methods that describe strategies for teaching content.

Some of the terminology used for the structures has been borrowed from work in intelligent tutoring systems (ITS). An ITS is an independent, computer-based system to support student learning. Researchers have been developing ITS for more than two decades (see, e.g., Sleeman & Brown, 1982). The ITS coordinates the use of test score information to inform formative decisions through the three components that comprise an ITS (Akhras & Self, 2002): a domain model, a teaching model, and a student model. The domain model is a structured representation of the knowledge and skills that constitute the construct. The teaching model represents the instructional methods used by the ITS in which instructional activities are selected and presented to the learner, and through which the learner's response is interpreted. The student model is an individualized representation of a student's current understanding of the

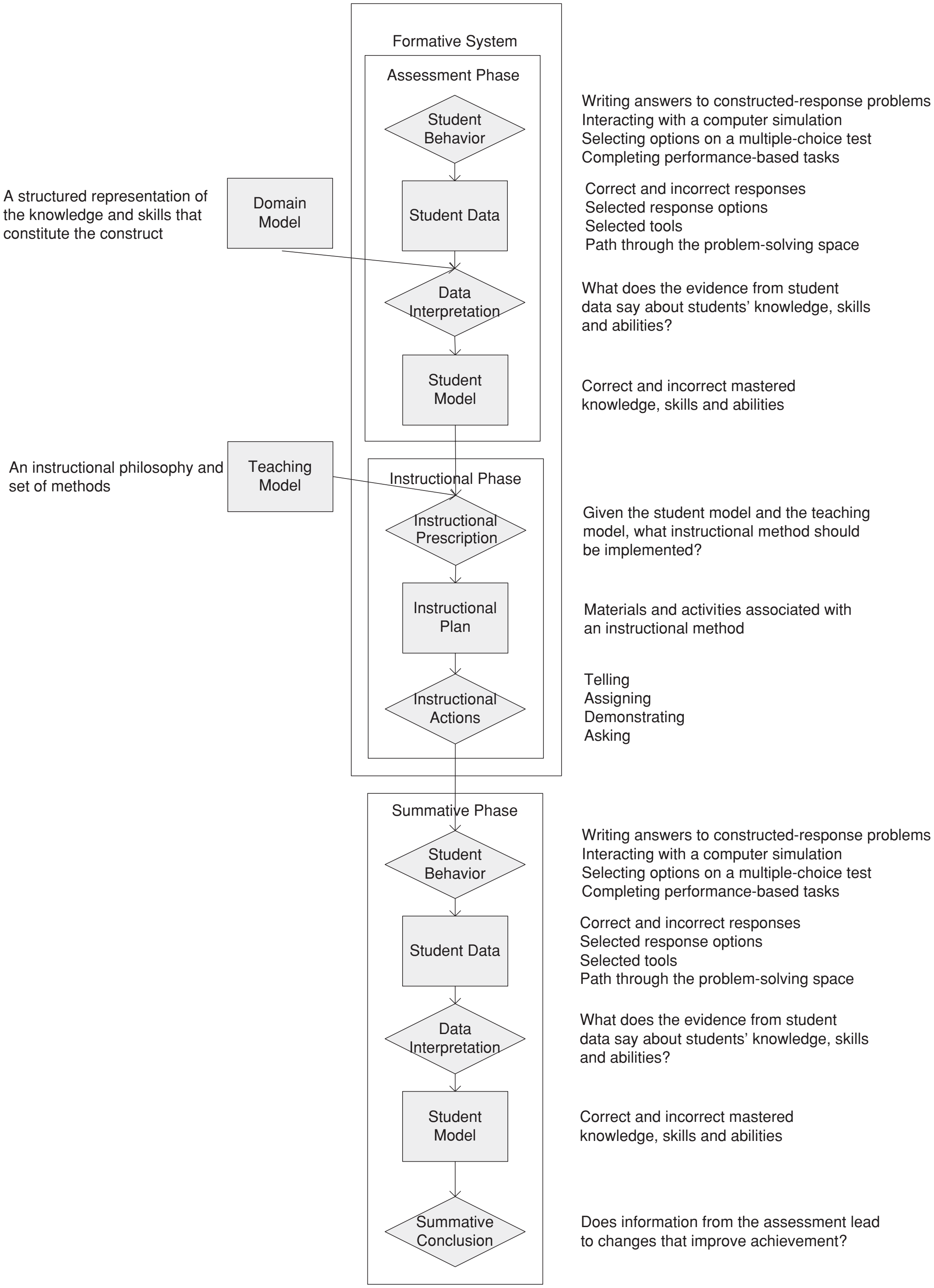

Formative System
Assessment Phase
Student Behavior
Writing answers to constructed-response problems
Interacting with a computer simulation
Selecting options on a multiple-choice test
Completing performance-based tasks
A structured representation of the knowledge and skills that constitute the construct
Domain Model
Student Data
Correct and incorrect responses
Selected response options
Selected tools
Path through the problem-solving space
Data Interpretation
What does the evidence from student data say about students' knowledge, skills and abilities?
Student Model
Correct and incorrect mastered knowledge, skills and abilities
An instructional philosophy and set of methods
Teaching Model
Instructional Phase
Instructional Prescription
Given the student model and the teaching model, what instructional method should be implemented?
Instructional Plan
Materials and activities associated with an instructional method
Instructional Actions
Telling
Assigning
Demonstrating
Asking
Summative Phase
Student Behavior
Writing answers to constructed-response problems
Interacting with a computer simulation
Selecting options on a multiple-choice test
Completing performance-based tasks
Student Data
Correct and incorrect responses
Selected response options
Selected tools
Path through the problem-solving space
Data Interpretation
What does the evidence from student data say about students' knowledge, skills and abilities?
Student Model
Correct and incorrect mastered knowledge, skills and abilities
Summative Conclusion
Does information from the assessment lead to changes that improve achievement?

FIGURE 1. A general framework for evaluating the validity of formative claims.

subject matter that is inferred from observable behavior (VanLehn, 1998). The framework for evaluating formative claims includes all three of these components as structures.

The actions are represented in the figure as diamond shapes. The actions may be interpretations made using structures. A sequence of interpretations is required to move from initial student behavior to a summative conclusion that the gap has been narrowed between students' current state of achievement and the targeted state of achievement. For example, interpretations are made of students' scores using both the domain model and the student data. Alternatively, actions may be implementations of interpretations. For example, instruction is implemented based on interpretations made using the teaching model and the student model.

Assessment Phase

The assessment phase includes the following three structures:

- Student data. The student data structure represents a subset of student behavior. An assessment does not capture all student behavior. Rather, the assessment captures selected aspects of student behavior such as the sequence of correct and incorrect responses, the response options selected, the manner that tools are manipulated, or a path through a problem-solving space. These aspects of behavior are captured but other aspects of behavior are ignored.
- Domain model. The domain model is the conceptualization of the domain, what Messick (1989) refers to as the construct theory. The domain model represents both the correct and incorrect knowledge and skills that constitute the construct. This includes the knowledge and skills to be transferred to the learner during instruction. The domain model likely includes far more than just a list of facts. For a classroom teacher,

> Teachers commonly have a far more extensive and elaborate knowledge base than their students. This includes straightforward factual matters (for example, the author of a particular book, or the planet nearest the sun, or the common form of the normal equations in simple linear regression) that enable them to recognize immediately whether a particular student's response is correct, partially correct, or incorrect, or whether the idea of correctness makes any sense in the context. It also includes procedural knowledge (for example, the variety of ways to do something, and which ones are better than others) and what might be termed a connoisseur's knowledge of a field or discipline (Sadler, 1998).

- Student model. The student model is a dynamic representation of the correct and incorrect knowledge, skills, and abilities the student has likely mastered. This individualized representation of a student's current understanding of the subject matter is inferred from observable behavior. The student model includes three of the four elements that Black and Wiliam (1998) require for a feedback system:

(1) Information on the actual level of some measurable attribute;
(2) Information on the reference level of that attribute;
(3) A mechanism for comparing the two levels, and generating information about the gap between the two levels.

The fourth element that Black and Wiliam (1998) require for a feedback system, a mechanism by which the information can be used to alter the gap, is included in the instructional phrase in which assessment information is employed.

The assessment phase includes two actions. The first action, student behavior, is broadly conceived of as a goal-directed human activity to be pursued in a specified manner, context, or circumstances (Haertel & Wiley, 1993). The second action, data interpretation, is a generalization of test score interpretation that is the foundation of validity arguments. Data interpretation involves reasoning from a handful of particular things students say, do, or make in particular circumstances, to their status on more broadly construed knowledge, skills, and abilities that constitute the student model.

Data interpretation may involve an intuitive process, a statistical algorithm, and many variations between these two extremes. As an example of an intuitive process, Wiliam and Black (1996) explain that the classroom teacher elicits and examines evidence of attainment based on an internal model of what it is to "understand" the ideas in question, trying to establish whether students share this model. "Provided the students' answers are consistent with the teacher's model, they will be regarded as having understood the topic" (Wiliam & Black, 1996, p. 543). As an example of a statistical algorithm, item response theory is used to aggregate students' responses so as to estimate students' ability levels. Alternative statistical algorithms have been proposed to estimate students' status relative to information processing components such as strategy use and use of misconceptions (Fu & Li, 2007).

Data interpretation is value laden. The role of values in interpreting student responses has been recognized by Tittle (1994) who argues that when a teacher questions a student, the teacher's beliefs will influence both the questions asked and the way that answers are interpreted. But values also play a role in interpreting student responses using statistical algorithms. For example, the application of the Rasch model (Rasch, 1960) in scoring and scaling student responses implicitly accepts that a single dimension sufficiently summarizes student status.

Instructional Phase

The instructional phrase is that part of the system within which assessment information is employed. The instructional phase includes two structures: the teaching model and the instructional plan. The first structure, the teaching model, consists of the available instructional methods and the educational philosophy used to select the instructional materials and activities presented to the learner. Assumptions about the psychology of learning, including assumptions about the motivations and self-perceptions of students, may be implicit or explicit in the teaching model. The second structure, the instructional plan, is the set of methods and materials intended to be used with the student. The instructional plan is the result of the instructional prescription.

The instructional phase includes two actions. First, the instructional prescription is the selection of the instructional methods and materials to use with the student. The selection of instructional methods is based on the coordination of the student model and the teaching model. The instructional prescription is the mechanism used to alter the gap between a student's current understanding and the targeted understanding. Second, the instructional action is the implementation of the activities and materials associated with the selected instructional plans. Note that the successful instructional action requires the motivated cooperation of the student if the system does or doesn't involve self-assessment, recognizing that the students are at once the agents of change and the thing that is changed (Brookhart, 2003; Stiggins, 2005).

Summative Phase

In the last phase, the summative phase, information from student behavior is used as evidence to reach a conclusion with regard to the formative claim of the initial phase. The components within the summative phase mirror the components within the formative phase. This duplication of components reflects the importance of the function the assessment information plays rather than the nature of the assessment in distinguishing formative from summative assessment.

The summative phase includes two structures: the student data and the student model. Both the student data and the student model structures are defined as they were within the formative phase. The nature of the information captured in those structures is unchanged but the nature of the inferences made using that information does change between the formative and summative phases.

The summative phase also includes three actions:

- Student behavior. Student behavior is defined just as within the formative phase.
- Data interpretation. Data interpretation is concerned with examining what the evidence from student data says about students' knowledge, skills, and abilities following instruction. This is the same interpretation attempted before instruction was given. The difference between the earlier interpretation and the current interpretation is in the time relative to instruction that the interpretation is attempted but not in the nature of the arguments and evidence that are relevant.
- Summative conclusion. The last interpretation concerns the summative conclusion drawn with regard to the initial assessment. Within this context, can assessment information be used to narrow the achievement gap? The validity of the three preceding interpretations must be supported before any summative conclusion can be entertained on the formative nature of the initial assessment information.

Validity Claim

The claim that information from assessment scores may be used to improve student achievement, like any validity claim, must be appraised using both empirical evidence and reasoned arguments (Kane, 2006; Messick, 1989). This claim of formative interpretation and use is evaluated in the same manner as test score interpretation and use more generally. Following the framework for evaluating the formative claim, the interpretations of assessment information are evaluated in a cascading fashion with later interpretations incorporating earlier interpretations. We argue that this validity evaluation process applies equally well to student self-assessment, integrated classroom activities, formative evaluations of multi-year curricular reforms, and more traditional testing.

An assessment system has the potential to provide formative information not because of the rapidity of feedback or because any particular individual is responsible for changes. The crucial factors that identify an assessment as a potential source of formative information include the availability of evidence of performances (current state of achievement), that the evidence is meaningfully related to the relevant criterion (target state of achievement), and that interpretations of the evidence can be used to make changes that effect progress toward the criterion (Stiggins, 2001; Wiliam, 2006; Wiliam & Black, 1996). In the next section, we illustrate the evaluation of a formative claim within a system through the application of our evaluative framework.

Example

This section presents an example selected to illustrate the framework for evaluating the validity of formative claims. This example describes the use of cognitive diagnostic assessment to tutor the ability to solve linear equations. Evaluation of the formative claim uses experimental design and random assignment to treatment. This example was selected for two reasons. First, to illustrate the difficulty of causally linking the use of information provided by an assessment to students' achievement gains. Second, to illustrate the systemic relationship between the validity concepts of score interpretation and score use. In this example, reasonable data interpretation in the assessment phase fails to lead to effective instructional prescription in the instructional phase.

Tutoring Algebra

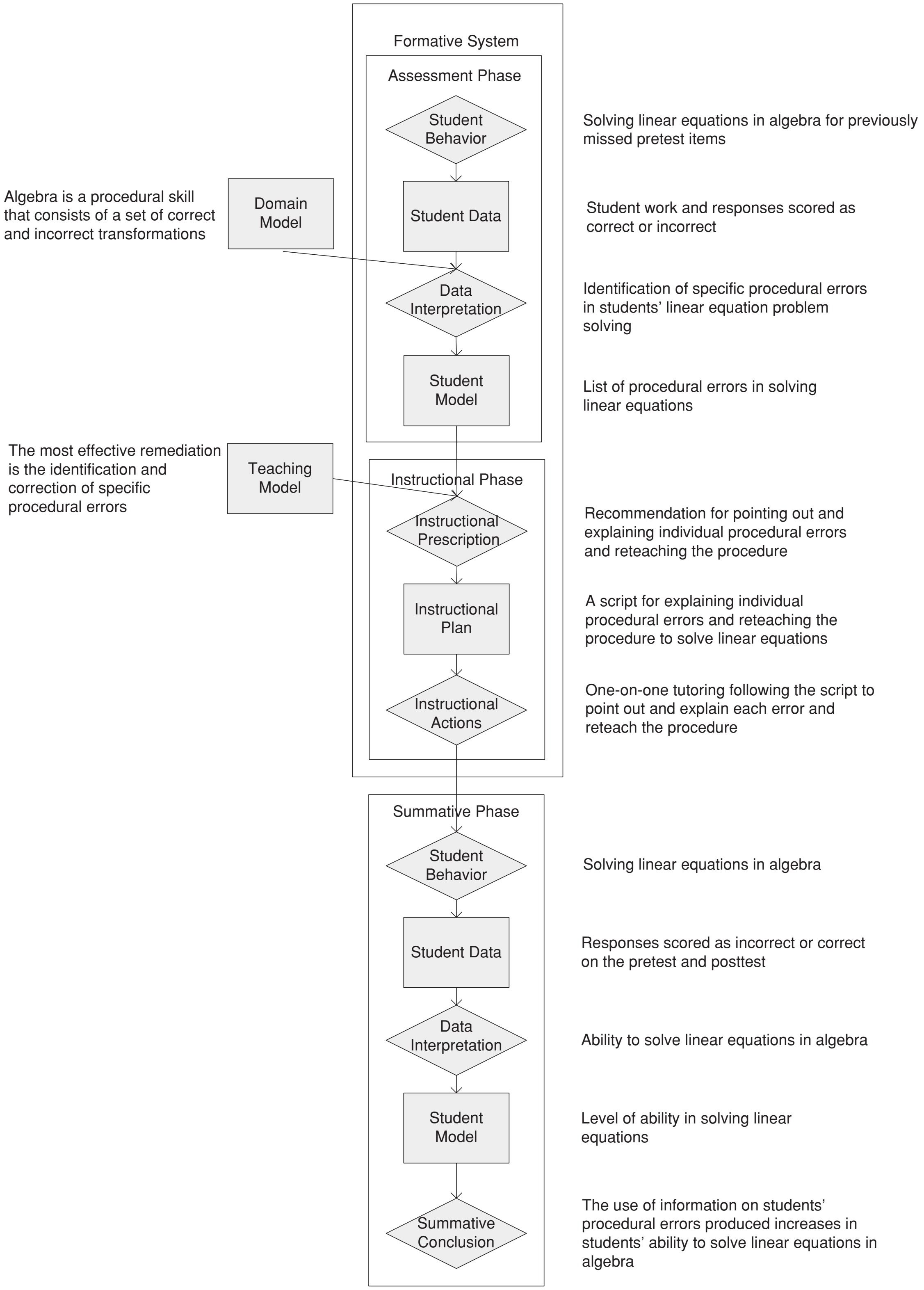
The example is provided by an experimental study in tutoring the ability to solve linear equations in algebra reported by Sleeman, Kelly, Martinak, Ward, and Moore (1989). For this example, the formative claim is that tutors' use of information about students' procedural errors in solving linear equations to tailor feedback narrowed the gap between students' current linear equation solving level and their targeted level. Sleeman et al. (1989) designed several experiments to test this causal claim. The first of these studies will be used to illustrate how the evaluation framework may be used to organize empirical evidence and reasoned arguments in support of the formative claim.

In the first study reported by Sleeman et al. (1989), students were recruited from two second- and third-year mathematics classes (students were approximately 13 or 14 years old) in a school in Scotland where they had received algebra instruction that was largely procedural in nature; algebra was treated as a series of transformations without extensive reference to possible meaning. The students were administered a 20-item pretest and those students who answered fewer than 16 items correct on the pretest were recruited for the study. Following tutoring, the students were administered a 20-item posttest constructed to be parallel in content and difficulty to the pretest. Both the pretest and posttest consisted of solving single algebra equations containing one unknown.

Students were randomly assigned to one of the two tutoring conditions in the study. Across both conditions, each student received approximately 35 minutes of individual tutoring designed to improve their procedural skills in solving linear equations. During tutoring, students first reworked an item incorrectly answered on the pretest. If the item was again answered incorrectly, the student received tutoring. The student then attempted at most two more items equivalent in format and difficulty to the first item and received additional tutoring if either item was answered incorrectly. Each student reworked all of the items that he or she had answered incorrectly on the pretest. In the first condition, the tutor pointed out and then explained the procedural error the student had made on the problem. The tutor then retaught the correct procedure. In the second condition, the tutor simply retaught the correct procedure.

Students showed statistically significant gains in the number of correct items from pretest to posttest. However, students in the second condition showed the same gains as students in the first condition. Students who were tutored using specific error remediation as well as reteaching showed the same gains as students who were tutored using only reteaching.

Each condition in the Sleeman et al. (1989) study may be considered an individual formative system. Both formative systems share a number of components including the same student data in both the assessment phase and the summative phase and the same data interpretation in the summative phase. However, the first condition, with the series of interpretations shown in Figure 2, may be considered a formative system



Formative System
Assessment Phase
Student Behavior
Solving linear equations in algebra for previously missed pretest items
Algebra is a procedural skill that consists of a set of correct and incorrect transformations
Domain Model
Student Data
Student work and responses scored as correct or incorrect
Data Interpretation
Identification of specific procedural errors in students' linear equation problem solving
Student Model
List of procedural errors in solving linear equations
The most effective remediation is the identification and correction of specific procedural errors
Teaching Model
Instructional Phase
Instructional Prescription
Recommendation for pointing out and explaining individual procedural errors and reteaching the procedure
Instructional Plan
A script for explaining individual procedural errors and reteaching the procedure to solve linear equations
Instructional Actions
One-on-one tutoring following the script to point out and explain each error and reteach the procedure
Summative Phase
Student Behavior
Solving linear equations in algebra
Student Data
Responses scored as incorrect or correct on the pretest and posttest
Data Interpretation
Ability to solve linear equations in algebra
Student Model
Level of ability in solving linear equations
Summative Conclusion
The use of information on students' procedural errors produced increases in students' ability to solve linear equations in algebra

FIGURE 2. A framework for a formative system using the identification of specific procedural errors.

in which the assessment information used by the tutor was the identification of specific procedural errors and the instruction was remediation of these errors. In contrast, the second condition, with the series of interpretations shown in Figure 3, may be considered a formative system in which the assessment information used by the tutor was the identification of incorrect procedures and the instruction was reteaching the correct procedure.

The conclusion could be reasonably drawn that the treatment, the tutoring using information about students' problem solving, caused students' gains in ability to solve linear equations. Despite the lack of a control group, threats

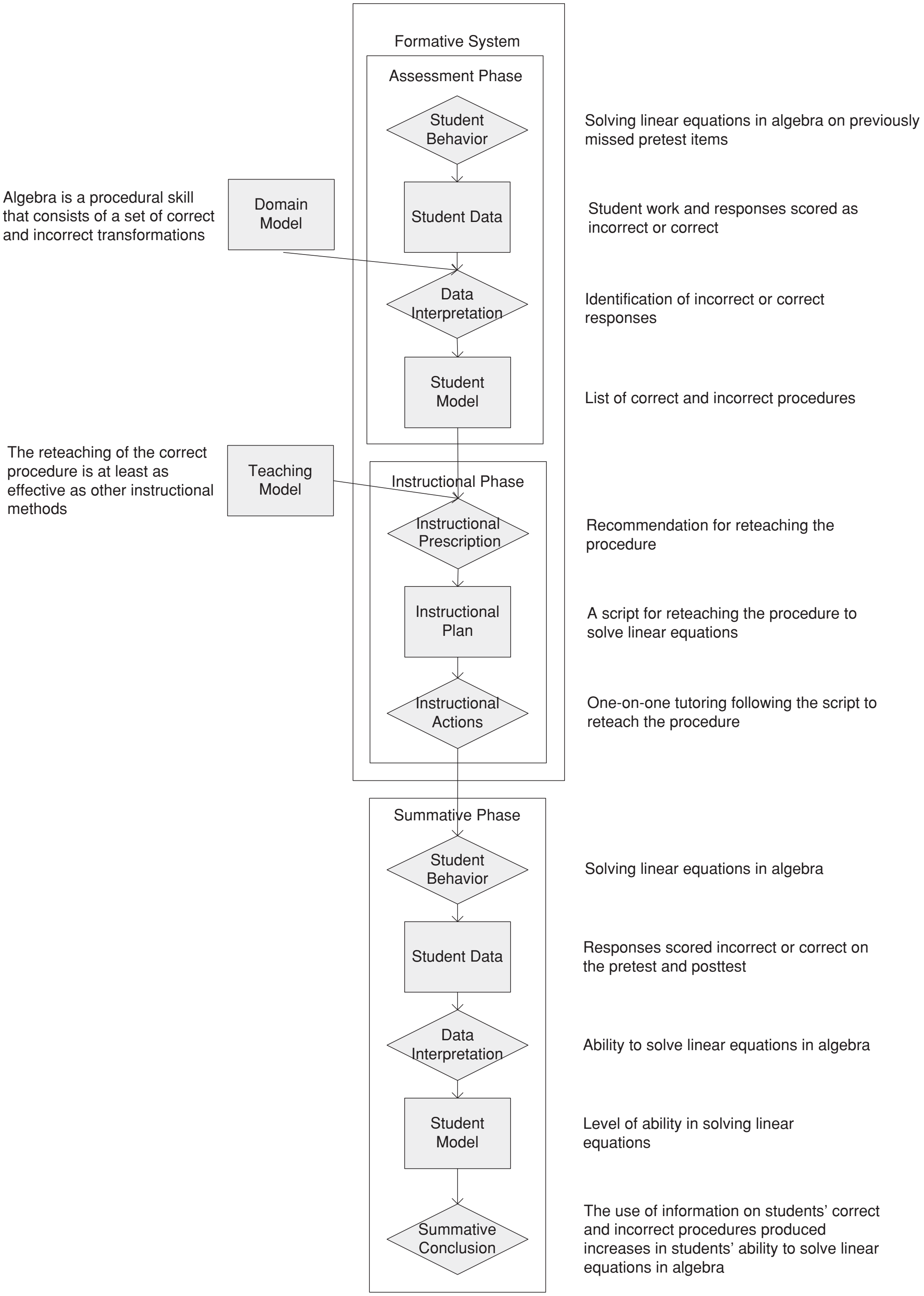

Formative System
Assessment Phase
Student Behavior
Solving linear equations in algebra on previously missed pretest items
Algebra is a procedural skill that consists of a set of correct and incorrect transformations
Domain Model
Student Data
Student work and responses scored as incorrect or correct
Data Interpretation
Identification of incorrect or correct responses
Student Model
List of correct and incorrect procedures
The reteaching of the correct procedure is at least as effective as other instructional methods
Teaching Model
Instructional Phase
Instructional Prescription
Recommendation for reteaching the procedure
Instructional Plan
A script for reteaching the procedure to solve linear equations
Instructional Actions
One-on-one tutoring following the script to reteach the procedure
Summative Phase
Student Behavior
Solving linear equations in algebra
Student Data
Responses scored incorrect or correct on the pretest and posttest
Data Interpretation
Ability to solve linear equations in algebra
Student Model
Level of ability in solving linear equations
Summative Conclusion
The use of information on students' correct and incorrect procedures produced increases in students' ability to solve linear equations in algebra

FIGURE 3. A framework for a formative system using reteaching.

to internal validity may be discounted by arguing that the students' knowledge of linear algebra was low and stable and students were unlikely to learn linear algebra from their homes, friends, or other academic courses (Shadish, Cook, & Campbell, 2002). But evidence of student gains that coincide with tutoring solving linear equations is not enough to causally link tutors' use of information about students' procedural errors to gains in student learning. The gains from pretest to posttest provide prima facie reason to investigate tutors' use of information about students' procedural errors. However, for the use of information about students' problem solving to be appropriately labeled "formative," both empirical evidence and reasoned arguments must be offered supporting the series of cascading interpretations that comprise the system of tutoring linear equation solving. In the following section, we show how evidence from the Sleeman et al. (1989) study, along with reasoned arguments, supports one condition but not the other as a formative system.

Applying the Framework

The systemic framework for evaluating a formative claim, described in Figure 1, can aid the marshaling of evidence and the constructing of arguments to support a formative claim for the system of tutoring linear equation solving in algebra. The system of tutoring linear equation solving in algebra comprises a series of cascading interpretations with later interpretations incorporating earlier interpretations. Each of these interpretations must be supported before support can be provided for the formative claim that use of information on students' procedural errors improved students' ability to solve linear equations. Early interpretations in the system of tutoring linear equation solving must receive support before later interpretations because they are incorporated into later interpretations. In this section, we will use the evaluative framework and work through each interpretation starting at the beginning of the series. We use evidence from the Sleeman et al. (1989) study, along with reasoned arguments, to support or challenge each interpretation.

The first interpretation, part of the assessment phase, that requires support before support can be given to the formative system is data interpretation, or, in the context of this study, the interpretation of students' problem-solving behavior in solving linear equations. This interpretation is part of the evaluative framework for condition 1 in Figure 2 and for condition 2 in Figure 3. As a comparison of Figures 2 and 3 shows, the same student behavior and student data are available in both condition 1 and condition 2. But the tutors in the two different conditions interpret the student work and scored responses differently. Under condition 1, student work and scored responses are interpreted as specific procedural errors. Under condition 2, student work and scored responses are interpreted as correct or incorrect procedures.

Support for condition 1 in which tutors interpret students' problem-solving behavior as procedural errors is provided in Sleeman et al. (1989) by an extensive literature review used to identify potential errors. The study authors reviewed psychological research on algebra problem-solving spanning more than 55 years that attempted to identify individual errors. This support might be classified by Kane (2006) as evaluating the backing for a theory-based inference.

Support for condition 2 in which tutors interpret students' problem-solving behavior as the application of correct or incorrect procedures rests on the expertise of the tutors. Tutors were university researchers who have studied algebra tutoring extensively. These judgments by qualified teachers are reasonable to accept at face value (Kane, 2006).

The second interpretation, part of the instructional phase, is instructional prescription, or, in the context of this study, the interpretation of the student and teaching models to recommend an instructional plan. A comparison of Figures 2 and 3 reveals that different teaching models and student models are used in the two conditions to make different recommendations for an instructional plan. Under condition 1, the teaching model and student model are used to recommend pointing out and explaining individual procedural errors and then reteaching the procedure. Under condition 2, the teaching model and student model are used to recommend reteaching the procedure.

Sleeman et al. (1989) offer support for both instructional prescriptions. They cite reviews of mathematics learning (Brown & Burton, 1978; Resnick, 1984) and empirical studies (Swan, 1983) that conclude error-specific remediation is superior to reteaching. But they also cite a number of studies of teacher behavior that found teachers generally do not diagnose specific errors (Kelly & Sleeman, 1986; Martinak, Schneider, & Sleeman, 1987; Putnam, 1987). This research found that teachers favored a review of the curriculum.

The third interpretation, part of the summative phase, that must be supported with evidence and arguments before a summative claim can be supported is data interpretation, or, in the context of this study, the interpretation of pretest and posttest scores as representing students' ability to solve linear equations. This interpretation is identical across condition 1 and condition 2. In the Sleeman et al. (1989) study, support for interpreting pretest and posttest scores as representing students' ability to solve linear equations would come from examining the content of the pretest and posttest.

The final interpretation that must be supported with evidence and arguments before a formative claim can be made is the summative conclusion, or, in the context of the Sleeman et al. (1989) study, that gains from pretest to posttest reflect the use of information about students' problem solving. Students in both conditions showed statistically significant gains in number correct from pretest to posttest. But students in the second condition showed the same gains as students in the first condition. Students who were tutored using specific error remediation as well as reteaching showed the same gains as students who were tutored using only reteaching.

If each condition is considered in isolation, the use of assessment information might be considered formative. For condition 1, the conclusion could be reasonable that tutoring using information about students' procedural errors caused students' gains in ability to solve linear equations. For condition 2, a similar conclusion may be reasonable that tutoring using information about students' incorrect or correct procedures caused students' gains in ability to solve linear equations.

But taken together, the evidence indicates that tutors' use of the additional information about students' procedural errors caused no gain in students' ability to solve linear equations. The formative claim appears reasonable for condition 2 but not for condition 1. The fault lies not in the interpretation of the student data but in the prescription based on the teaching model.

Implications

In this section, we explore three implications that may be inferred given this framework that represents formative assessment information as a component of a system of coordinated assessment and instruction that eventually leads to improvements in student achievement. First, this framework implies that assessment information is likely to be used effectively to improve student achievement when the information fits as a component of a system of explicitly coordinated assessment and instruction. Second, this framework implies that validity evidence supporting the conclusion that any particular set of assessment results as formative holds only within a limited context. Finally, this framework implies that test developers of assessments that are intended to offer information that can be used to improve student achievement should consider the series of interpretive arguments in designing their assessment.

Systemic Argument

Explicit in our proposal of this evaluative framework is the claim that for assessment information to serve a formative function, the information must fit as a component of a system of coordinated assessment and instruction. The study by Sleeman et al. (1989) offers an illustration of the systemic nature of the formative claim. In that study, the interpretation of student work and scored responses as specific procedural errors is a reasonable data interpretation. Yet that assessment information fails to serve a formative function because the instructional prescription is not effective. Each component of the formative system must work for the assessment information to serve a formative function.

The need for coordination is further illustrated by the example of the Sleeman et al. (1989) study in which the data interpretation used to build the student model should be coordinated with the teaching model. In the Sleeman et al. (1989) study, student work and scored responses need not be interpreted as specific procedural errors if the only available teaching model is to reteach the lesson.

Contextual Claim

The need for a systemic framework to evaluate a formative claim demonstrates that assessments used for formative purposes are inextricably linked with the surrounding context. Any discussion or evaluation of the assessment itself cannot be appropriately done in isolation. Context is an essential element in building an argument that a given assessment can serve a formative purpose. An assessment can successfully serve a formative purpose in one scenario, but fail to serve the intended purpose in another scenario. The generalizability of the formative function of assessment information is what Shadish et al. (2002) refer to as external validity.

For example, Sleeman et al. (1989) found that using the assessment to identify specific procedural errors offered no more formative information beyond that offered by using the assessment to identify incorrect and correct procedures. Does the formative claim to identify incorrect and correct procedures extend beyond solving linear equations in algebra? Does the claim extend across mathematics? Across other domains? Establishing the boundaries of a formative claim is a greater challenge than simply evaluating a claim.

The boundary of the formative claim is established through evidence and argument. The boundary is likely to be established along a number of dimensions including the characteristics of the learner, the characteristics of instruction and the nature of the targeted knowledge and skills. The learner may vary in level of achievement, locus of control, age, or other variables. Instruction may vary in characteristics such as self-directed or other-directed or lecture or discovery style. The targeted knowledge and skills may vary in such characteristics as procedural or conceptual or content domain.

Prescriptive Use

Rather than using this systemic framework reactively to evaluate formative claims after assessments have been developed, we advocate using the systemic framework in a proactive manner. We encourage test developers, and others developing formative practices, to use the framework to prescribe the parameters of the formative system during the design phase. This includes making clear statements about the inferences to be drawn based on the initial student behavior, the prescription of instructional activities within a teaching model, the inferences to be drawn based on student behavior following the application of the instructional activities, and the overall interpretation of the system. This move from evaluation to prescription is based on Kane's (1992, 2006) argument-based approach to validity investigation that relies on the proposed interpretations and uses of test scores being clearly specified during test development.

The prescriptive use of this systemic framework places greater emphasis on supporting the consumers of assessment information (e.g., educators, parents, and students). An application of an assessment for formative purposes must occur in concert with a sustained support effort that facilitates that application in the classroom, in the home, or wherever learning occurs. As Wiliam wrote in an introduction to a 2006 volume of *Educational Assessment* dedicated to the topic of formative assessment, "In other words, the task of improving formative assessment is substantially, if not mainly, about teacher professional development" (p. 287).

Finally, we conclude the article by arguing that developers who intend information from their assessments to serve a formative purpose should anticipate the need for a validity argument. In anticipation of this requirement, these assessment developers should make explicit the links between test score information from their assessment to the selection of instructional actions whose implementation leads to gains in student learning. We suggest that assessment developers who intend their assessments to be formative should conceive, in contrast to a stand-alone assessment, a system of coordinated assessment and instruction.

Conclusion

The labeling of a particular test as a formative assessment implies an overall evaluative judgment of the inferences made and the actions taken by educators based on assessment information. As we argue here, the term formative assessment implies that the assessment results, whether from classroom interrogation strategies or standardized assessments, can be used to make changes. Referring to a specific assessment as a formative test is as misguided as referring to a specific assessment as a valid test. A clearer description would refer to the formative use of test score information. Just as when the phrase "test validity" is used as shorthand for score interpretation and use, use of the phrase "formative assessment" in technical discourse should be understood as shorthand for formative uses of test score information. Of course, outside of technical discourse, reference to an assessment as a formative assessment is suitable for informal conversation.

Furthermore, the prescriptive use of the evaluative framework proposed here emphasizes the importance of the instruction that is implemented based on assessment information. Formative assessment only has value if the instruction results in increased positive outcomes relative to the expected outcome if no changes had been made. Different instructional changes or interventions are likely to result in different outcomes, and some changes, no matter how well intentioned or planned, may not have a positive impact. The topic of what changes should be implemented following the initial interpretation of student performance, how those changes should be implemented, and how they should be supported is crucial to ensuring that formative assessments fulfill their promise to improve outcomes.

References

Akhras, F. N., & Self, J. A. (2002). Beyond intelligent tutoring systems: Situations, interactions, processes and affordances. *Instructional Science, 30*, 1–30.

Baker, E. L. (2004). *Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform*. CSE Report 645. Los Angeles, CA: Center for the Study of Evaluation (CSE)/National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Educational Assessment: Principles, Policy and Practice, 5*, 7–74.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practices, 22*(4), 4–54.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155–192.

Coombs, C. H., Raiffa, H., & Thrall, R. M. (1954). Some views on mathematical models and measurement theory. *Psychological Review, 61*, 132–144.

Fu, J., & Li, Y. (2007, April). *Cognitively diagnostic psychometric models: An integrative review*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17*(2), 16–19, 34.

Haertel, E., & Wiley, D. (1993). Representations of ability structures: Implications for testing. In N. Fredrickson, R. Mislevy & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359–384). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kelly, A. E., & Sleeman, D. (1986). *A study of diagnostic and remedial techniques used by master algebra teachers*. Tech. Rep. AUCSiTR8708. Aberdeen, Scotland: University of Aberdeen, Department of Computer Science.

Martinak, R., Schneider, B., & Sleeman, D. (1987, April). *A comparative analysis of approaches for correcting algebra errors via an intelligent tutoring system*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9–20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*, 13–48.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–16.

Resnick, L. (1984). *Beyond error analysis: The role of understanding in elementary school arithmetic*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, 5, 77–84.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sleeman, D. H., & Brown, J. S. (Eds.) (1982). *Intelligent tutoring systems*. New York: Academic Press.

Sleeman, D., Kelly, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science, 13*, 551–568.

Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice, 20*(1), 5–15.

Stiggins, R. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324–328.

Swan, M. B. (1983). *Teaching decimal place value: A comparative study of conflict and positively-only approaches*. Research Rep. No. 31. Nottingham, UK: University of Nottingham, Sheel Center for Mathematical Education.

Tittle, C. K. (1994). Toward an educational psychology of assessment for teaching and learning: Theories, contexts, and validation arguments. *Educational Psychologist, 29*, 149–162.

VanLehn, K. (1998). Analogy events: How examples are used during problem solving. *Cognitive Science, 22*(3), 347–388.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment, 11*(3 & 4), 283–289.

Wiliam, D., & Black, P. (1996) Meaning and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537–548.